Sum of Squares & R Squared

Ian He

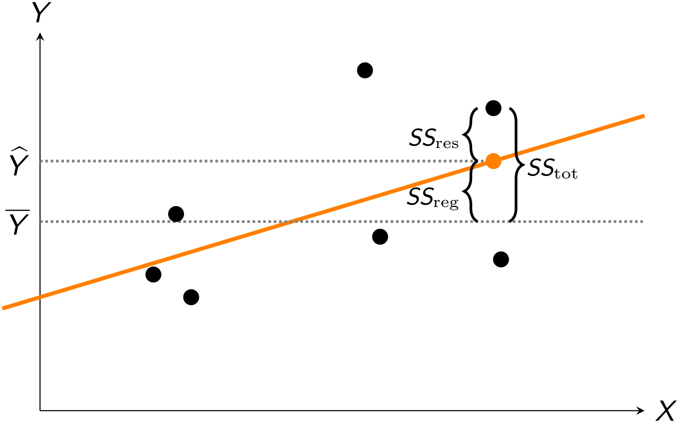
Amateur Explorer of $\mathcal{E}\text{con}\phi\text{metric}\$$

September 22, 2023

Table of Contents

Overview

The **sum of squares** is a statistical measure of **variability**. We usually decompose variability into three types of sum of squares, as shown below.



Total Sum of Squares

The **total sum of squares** (TSS) or **sum of squares total** (SST or SS_{tot}) is the sum of squared differences between the observed dependent variables and the overall mean.

$$SS_{\text{tot}} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

This is similar (not identical) to the sample variance of dependent variable in descriptive statistics.

Explained Sum of Squares

The **explained sum of squares** (ESS), also called **sum of squares due to regression** (SSR or SS_{reg}) or **model sum of squares** (MSS), is the sum of the differences between the fitted value and the mean of the dependent variable.

$$SS_{\text{reg}} = \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y} \right)^2$$

Residual Sum of Squares

The **residual sum of squares** (RSS), also called the **sum of squared residuals** (SSR or SS_{res}) or the **sum of squared estimate of errors** (SSE), is the sum of the squares of residuals (deviations fitted from actual empirical values of data).

$$SS_{\mathrm{res}} = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i \right)^2$$

Alert: Be careful when seeing the abbreviation SSR. There's no consensus on abbreviations of **sum of squares due to regression** and **sum of squared residuals**, so SSR could refer to either term in different texts.

Analysis-of-Variance formula

The following equality is generally true in the OLS regression:

$$SS_{\mathrm{tot}} = SS_{\mathrm{res}} + SS_{\mathrm{reg}}$$

or equivalently,

$$\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 + \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2$$

This equation is sometimes called the **analysis-of-variance formula** for the OLS regression. See Wikipedia for its proof.

Sequential Sum of Squares

The **sequential sum of squares**, also called the **extra sum of squares**, can be viewed in two ways:

- It is the reduction in the residual sum of squares (SS_{res}) when one or more explanatory variables are added to the model.
- It is the increase in the explained sum of squares (SS_{reg}) when one or more explanatory variables are added to the model.

A sequential sum of squares quantifies how much more variability we explain (i.e., the increase in SS_{reg}) or alternatively how much error we reduce (i.e., the reduction in SS_{res}).

Notation: $SS_{\mathrm{res}}(X_2|X_1)$ or $SS_{\mathrm{reg}}(X_2|X_1)$ denotes the sequential sum of squares obtained by adding X_2 to a model already including the explanatory variable X_1 and a constant term.

R Squared I

The **coefficient of determination**, denoted R^2 and pronounced "**R squared**", provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

Its most general definition is

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{SS_{\mathrm{res}}}{SS_{\mathrm{tot}}}$$

When the relation $SS_{\mathrm{tot}} = SS_{\mathrm{reg}} + SS_{\mathrm{res}}$ holds, the above definition is equivalent to

$$R^2 = \frac{SS_{\mathrm{reg}}}{SS_{\mathrm{tot}}} = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{\text{explained variance}}{\text{total variance}}$$

R Squared II

R^2 is a measure of the goodness of fit of a regression model, but it does **NOT** indicate whether

- omitted-variable bias exists;
- the most appropriate set of independent variables has been chosen;
- there is collinearity present in the data on the explanatory variables;
- the correct regression was used;
- the independent variables are a cause of the changes in the dependent variable;
- there are enough data points to make a solid conclusion.

What's worse, R^2 increases as the number of variables in the model increase!

Adjusted R Squared I

Due to the phenomenon that the R^2 is at least weakly increasing when extra explanatory variables are added to the model, the **adjusted** R^2 (denoted $\overline{R}^2$) was invented. The explanation of $\overline{R}^2$ is almost the same as R^2 but it penalizes the statistic as extra variables are included in the model.

There are many different ways of adjusting. A commonly used one is the correction proposed by Ezekiel (1930):

$$\overline{R}^2 = 1 - \frac{SS_{\text{res}} / \, df_{\text{res}}}{SS_{\text{tot}} / \, df_{\text{tot}}}$$

where df_{res} is the degrees of freedom of the estimate of the population variance around the model, and df_{tot} is the degrees of freedom of the estimate of the population variance around the mean.

Adjusted R Squared II

Assume that we have n observations and that k coefficients and an intercept are estimated. Then,

$$df_{\mathrm{res}} = n - k - 1$$
$$df_{\mathrm{tot}} = n - 1$$

Thus, the definition of $\overline{R}^2$ can be rewritten as

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

References

Ezekiel, M. J. B. (1930).
Methods of correlation analysis.
Wiley.

Hansen, B. E. (2022).
Econometrics.
Princeton University Press.